Nearest Neighbor Categorization for CASP Function Prediction

Karin Verspoor*, Judith Cohn†, Susan Mniszewski*, Cliff Joslyn*

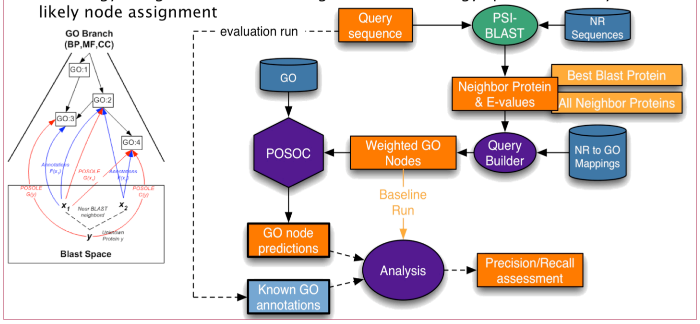
Computer and Computational Sciences Division, †Biological Division
Los Alamos National Laboratory

Motivation

We present the methods utilized in a system aimed at predicting the function of CASP protein targets, as represented by a node in the Gene Ontology. The strategy we follow is to (1) identify close neighbors of a target sequence in sequence space, (2) collect the Gene Ontology nodes associated with these neighbors in a curated data set (Swiss-Prot), and (3) categorize the collection of Gene Ontology nodes based on their distribution in the Gene Ontology structure, utilizing a technology called the POSOC, the POSet Ontology Categorizer. The resulting set of Gene Ontology nodes is interpreted as the most representative nodes for the function of the original target sequence.

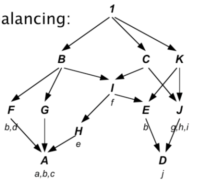
POSOLE: POSet Ontology Laboratory Environment

- a general environment for ontology experimentation
  - Graph representation of an ontology as a POSet
  - POSet statistics analysis (e.g. depth, width, average rank)
  - Algorithms for node categorization utilizing the structure of the ontology
- First Deployment: Ontology categorization for automated protein function annotation
  - Function: Gene Ontology node
  - Protein: target sequence or Swiss-Prot identifier
  - Map proteins to sets of potential Gene Ontology nodes
  - Ontology categorization: "clustering" nodes in ontology space to identify the most likely node assignment



POSOC: POSet Ontology Categorizer

- Given the Gene Ontology (GO) . . . And mappings to GO nodes . . .
- "Splatter" them over the GO . . . Where do they end up?
  - -- Concentrated? -- Dispersed?
  - -- Clustered? -- Overlapping or distinct?
  - -- High or low?
- Pseudo-distances between comparable nodes to measure vertical separation
- POSOC traverses the structure of the GO, percolating hits upwards, and calculating scores for GO nodes.
- Scores to rank-order nodes with respect to gene locations, balancing:
  - **Coverage:** Covering as many genes as possible
  - **Specificity:** But at the "lowest level" possible
- "Cluster" based on non-comparable high score nodes
- Example: Given labels (genes) *c, e, i . . .*
  - What node(s) *A,B, C, . . . ,K* are best to attend to?
    C, {*H*, *J*}, {*A*, *H*, *J*}
  - Depends on balance of specificity and coverage



CASP Evaluation Runs

- Goal: compare function predictions made by the system with known functions assigned to each input protein
- Test set: proteins with known Gene Ontology mappings from UniProt
  4530 SwissProt protein sequences derived from Protein Data Bank
- Runs

> **Baseline Best Blast**: Predictions are the GO nodes associated with non-identical protein scoring highest in the PSI-BLAST analysis. All predicted GO nodes are considered to be at rank 1.
>
> **Baseline Full Neighborhood**: Predictions are the GO nodes associated with ***all*** proteins matched in the PSI-BLAST analysis (with evalue < 10). The predictions are ranked according to the evalue of the corresponding PSI-BLAST match.
>
> **POSOC Best Blast**: Inputs to POSOC are the GO nodes associated with non-identical protein scoring highest in the PSI-BLAST analysis, weighted by evalue of the match.
>
> **POSOC Full Neighborhood**: Inputs to are the GO nodes associated with ***all*** proteins matched in the PSI-BLAST analysis, weighted by evalue of the match.

- Eliminate identity matches in PSI-BLAST from mappings used in prediction
  - Matches to protein with the same SwissProt Accession ID
  - Matches to protein with the same SwissProt Entry ID
  - Matches to protein with an e-value < 10^{-130} or e-value <= protein to itself

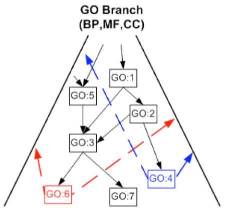
Hierarchical Evaluation Metrics

- Compare answers *F(x)* against predictions *G(x)*
- Precision/Recall
  - Precision = % of predictions that are correct

$$P = \frac{|F(x) \cap G(x)|}{|G(x)|}$$

  - Recall = % of known answers that are recovered

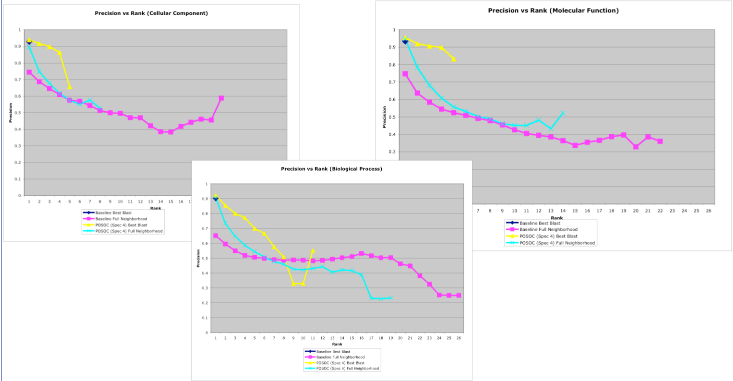
$$R = \frac{|F(x) \cap G(x)|}{|F(x)|}$$



- Extension to ranked list of predictions:
  Consider precision/recall at different ranks
- Extension to ontological predictions: when does a GO node *p* in *F(x)* count as a "match" against a *q* in *G(x)*?
  - What about siblings? Ancestors?
  - Adapt hierarchical precision/recall measure from Kiritchenko *et al* 2005

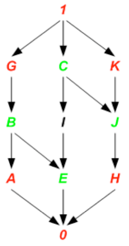
$$P = \sum_{q \in G(x)} \max_{p \in F(x)} \frac{|\uparrow p \cap \uparrow q|}{|\uparrow q|}$$

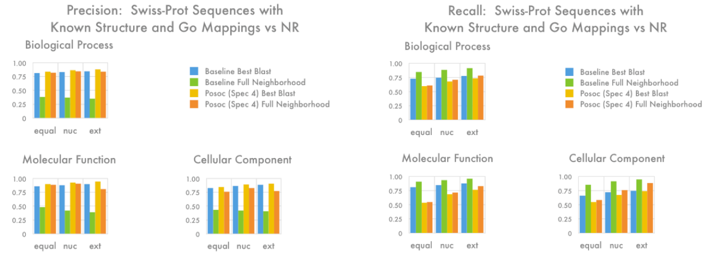
$$R = \sum_{p \in F(x)} \max_{q \in G(x)} \frac{|\uparrow p \cap \uparrow q|}{|\uparrow p|}$$



Ontology Distance Metrics

- How "far apart" are *p* and *q*?
- Genealogical approach:
  - Radius 0: **Equals**: Direct match
  - Radius 1: **Nuclear family**: Parents, children, siblings
  - Radius 2: **Extended family**: grandparents, grandchildren, cousins, aunts/uncles, nieces/nephews
- Towards a general formulation of metric-based poset distances and evaluation functions: under development





REFERENCES

- CA Joslyn: (2004) "Poset Ontologies and Concept Lattices as Semantic Hierarchies", in Conceptual Structures at Work, LNAI, v. 3127, ed. Wolff , pp. 287-302, Springer-Verlag, Berlin
- CA Joslyn and WJ Bruno: (2005) "Weighted Pseudo-Distances for Categorization in Semantic Hierarchies",Int. Conf. on Conceptual Structures, to appear in Lecture Notes on AI.
- CA Joslyn, SM Mniszewski, AW Fulmer and GG Heaton: (2004) "The Gene Ontology Categorizer", , v. 20:s1, BioInformatics, pp. 169-177
- S Kiritchenko, S Matwin, and AF Famili: (2005) "Functional Annotation of Genes Using Hierarchical Text Categorization", to appear in Proc. BioLINK SIG on Text Data Mining
- D Martin, M Berriman, and G Barton: (2004) "GOtcha: A New Method for Prediction of Protein Function Assessed by the Annotation of Seven Genomes", BMC Bioinformatics, 5:178
- D Pal and D Eisenberg, David: (2005) ``Inference of Protein Function from Protein Structure'', Structure, v. 13, pp. 121-130
- KM Verspoor, JD Cohn, SM Mniszewski, and CA Joslyn: (2004) "Nearest Neighbor Categorization for Function Prediction". In CASP 06 abstract book.
- KM Verspoor, JD Cohn, CA Joslyn, SM Mniszewski, A Rechtsteiner, LM Rocha, and T Simas: (2005) :"Protein Annotation as Term Categorization in the Gene Ontology Using Word Proximity Networks", BMC Bioinformatics, vol 6(suppl 1).